PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-184351
(43)Date of publication of application : 06.07.2001

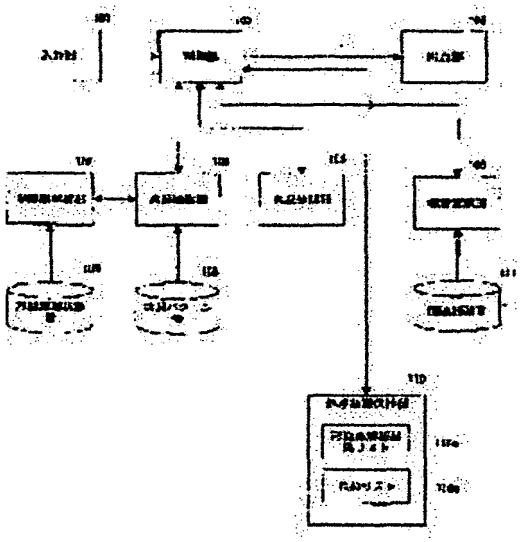
(51)Int.Cl. G06F 17/28
G06F 17/22
G06F 17/30

(21)Application number : 11-370936
(22)Date of filing : 27.12.1999

(71)Applicant : TOSHIBA CORP
(72)Inventor : ICHIMURA YUMI

(54) DOCUMENT INFORMATION EXTRACTING DEVICE AND DOCUMENT SORTING DEVICE

(57)Abstract:
PROBLEM TO BE SOLVED: To solve a problem that a means for a user to update a dictionary is not provided since the dictionary required for sorting documents is conventionally prepared in advance.
SOLUTION: While utilizing a concept definition dictionary in which a concept belonging to each of sorting axes an expression representing this concept are correlated for a plurality of preset plural sorting axes, concepts contained in each of documents are extracted and the document is sorted while using a compound concept created by combining concepts belonging to different sorting axes among the extracted concepts. Concerning such a system, whether it is necessary to update the dictionary is judged and when updating is required, the list of candidates to be registered in the dictionary is presented so that the document sorting system facilitated in the maintenance of the dictionary can be provided.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2001－184351
(P2001－184351A)
(43)公開日　平成13年７月６日(2001.7.6)

(51)Int.Cl.'	識別記号	ＦＩ		テーマコード'(参考)
Ｇ０６Ｆ　17/28		Ｇ０６Ｆ　15/38	Ｃ	５Ｂ００９
17/22		15/20	５２０Ｌ	５Ｂ０７５
17/30			５２２Ｌ	５Ｂ０９１
		15/40	３７０Ａ	
			３７０Ｊ	

審査請求　未請求　請求項の数６　ＯＬ　（全 18 頁）　最終頁に続く

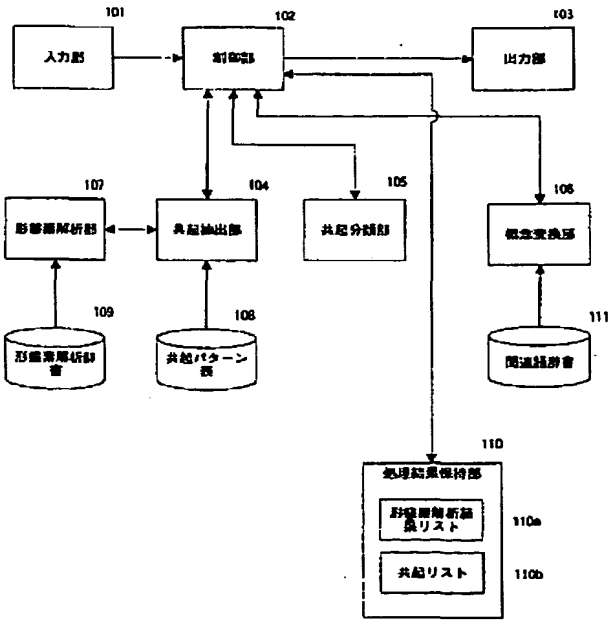
(21)出願番号　特願平11－370936
(22)出願日　平成11年12月27日(1999.12.27)

(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地
(72)発明者　市村　由美
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内
(74)代理人　100081732
弁理士　大胡　典夫　（外１名）

Ｆターム(参考)　5B009 MB22 ME13 ME22 MF03 VA02
5B075 ND03 NR12 PP02 PP03 PQ02
QP03 UU06
5B091 AA11 AB17 CC05

(54)【発明の名称】　文書情報抽出装置および文書分類装置

(57)【要約】
【課題】　従来、文書分類に必要な辞書はあらかじめ用意されているものであり、ユーザが辞書を更新する手段は提供されていなかった。
【解決手段】　あらかじめ設定された複数の分類軸に対して、各分類軸に属す概念と該概念を表す表現とを対応づけた概念定義辞書を利用して、各文書中に含まれる概念を抽出し、抽出された概念のうち、異なる分類軸に属す概念同士を組み合わせた複合概念を用いて文書を分類するシステムにおいて、辞書の更新が必要かどうかを判断し、更新の必要がある場合には辞書登録すべき候補の一覧を提示することにより、辞書のメンテナンスが容易な文書分類システムを提供する。



【特許請求の範囲】
【請求項１】　文書を入力する入力手段と、この入力手段により入力された文書を解析し、あらかじめ登録された共起パターンに適合する共起データを、少なくとも第一の構成要素，第２の構成要素，要素間の関係，第２の構成要素に付随する付属語列とともに記憶する共起抽出手段と、この共起抽出手段により抽出された共起データに関して、要素間の関係または第２の構成要素に付随する付属語列を参照して分類する共起分類手段とを具備することを特徴とする文書情報抽出装置。
【請求項２】　文書を入力する入力手段と、この入力手段により入力された文書を解析し、あらかじめ登録された共起パターンに適合する共起データを、少なくとも第一の構成要素，第２の構成要素，要素間の関係，第２の構成要素に付随する付属語列とともに記憶する共起抽出手段と、この共起抽出手段により抽出された共起データに関して、要素間の関係または第２の構成要素に付随する付属語列を参照して分類する共起分類手段と、同義語，類義語，シソーラス等の関連語を記憶した関連語辞書と、前記共起抽出手段により抽出された各共起データの構成要素を関連語辞書に記憶される関連語に置き換え、各共起データをより上位の概念へと変換する概念変換手段とを具備することを特徴とする文書情報抽出装置。
【請求項３】　複数の文書を入力する入力手段と、あらかじめ設定された複数の分類軸に対して、各分類軸に属す概念と該概念を表す表現とを対応づけた概念定義辞書と、この概念定義辞書を用いて、前記入力手段により入力された各文書中に含まれる概念を抽出する概念抽出手段と、この概念抽出手段により抽出された概念のうち、異なる分類軸に属す概念同士を組み合わせた複合概念を用いて前記入力された文書を分類する文書分類手段と、この文書分類手段の分類性能を評価する分類性能評価手段と、この分類性能評価手段の評価結果に基づき、前記概念定義辞書への辞書登録候補を自動生成する辞書作成手段と、この辞書作成手段により作成された辞書登録候補を表示する表示手段とを具備したことを特徴とする文書分類装置。
【請求項４】　前記分類性能評価手段は、各分類に属す文書件数とあらかじめ設定した値とを比較することにより性能評価を行うことを特徴とする請求項３記載の文書分類装置。
【請求項５】　前記分類性能評価手段は、入力された文書件数、いずれかの分類に属す文書件数、および、あらかじめ設定した値とを比較することにより性能評価を行うことを特徴とする請求項３記載の文書分類装置。
【請求項６】　前記辞書作成手段は、入力手段により入力された文書を解析し、あらかじめ登録した共起パターンに適合する共起データを、少なくとも、第１の構成要素、第２の構成要素、要素間の関係、第２の構成要素に付随する付属語列とともに記憶する共起抽出手段と、共起抽出手段により抽出された共起データに関して、要素間の関係または第２の構成要素に付随する付属語列を参照して、各共起データが属す分類軸を推定する分類軸推定手段と、　同義語、類義語、シソーラス等の関連語を記憶した関連語辞書と、共起抽出手段により抽出された各共起データの構成要素を関連語辞書に記憶された関連語に置き換え、各共起データをより上位の概念へと変換する概念変換手段とを有し、共起データ、分類軸、上位概念を対応づけて記憶することを特徴とする請求項３記載の文書分類装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、文章形式で記述されたデータを自動的に分類する文書分類システムにおいて利用される辞書を自動生成する文書情報抽出方法に関する。
【０００２】
【従来の技術】さまざまな場所で日々多様な情報が蓄積されているが、そのデータの８割以上が文章データであると言われている。このような文章データから情報を引き出すためには、基本的には人間が読むしかないのが現状である。
【０００３】たとえば、大量の営業報告書を分析することを考えると、目的に応じて必要なデータだけを形式を決めて記述しておけば、その分析は比較的容易である。しかし、そのような制限を設けると、設定された形式では記述できない情報は捨てられることになる。一方、営業活動における危機警告情報やノウハウ情報などをうまく捉える項目を予め用意することは困難である。そのため、報告書は文章記述を中心に作成されることになり、そのように作成された大量の文書から目的とする情報や知識を読み取るには、大変な労力が必要となる。
【０００４】その結果、管理者が業務上の意志決定をするために重要となる営業活動における危機警告情報やノウハウ情報は大量の書類の中に埋もれてしまい、営業報告書が十分に活用されているとはいえないのが現状である。
【０００５】上記問題点を解決する手段として、第５７回情報処理学会全国大会３－７５および３－７６にて、テキストマイニング技術によって文書を分析する方法が提案されている。
【０００６】この方法では、意図を含んだ情報を抽出するために、モダリティと構文情報を利用している。まず、あらかじめ「要望」「質問」「好評」「問題」といったモダリティに対応する評価ラベルを用意する。そして、構文情報として係り受け関係を想定した２語組を抽出し、重要語ラベルと述語・評価ラベルとの組合せパターンを用意して、情報抽出を行う。すなわち、重要語ラベルと呼ぶ名詞表現と評価ラベルを含んだ述語表現との

組合せによって、情報抽出を行っている。

【０００７】この方法によると、抽出される概念は、「○が□した」（ただし、□は評価ラベルつき）という複合概念である。つまり、抽出される内容は、「○を□したい」「○を□したくない」「○は○であった」といったような、いわば断片的な情報である。

【０００８】一方、業務上の意志決定に有効な情報とは、例えば「プロジェクトの進行状況はどうなっているか？」「あるアクションの結果はどうなっているのか？」「ある事例に遭遇した場合どういう行動をとればよいのか？」といったことであり、そのためにはさらに高度なテキストマイニング技術が求められる。

【０００９】上記問題を解決する手段として、当社先願発明である「業務支援システム」（市村、中山）において、アクションと結果を含む複数の分類軸に属す概念とその表現とを定義した概念定義辞書を利用して、各文書中に含まれる概念を抽出し、異なる分類軸に属す概念同士を組み合わせた複合概念を用いて文書を分析する方法が提案されている。しかし、前記特許においては、概念定義辞書はあらかじめ用意されているものであり、ユーザがその辞書を更新する手段は提供されていなかった。

【００１０】一方、文書から共起データを抽出する方法が、特許第２９４３４４７号「テキスト情報抽出装置とテキスト類似照合装置とテキスト検索システムとテキスト情報抽出方法とテキスト類似照合方法、及び、質問解析装置」に開示されている。この方法によると、与えられた文書を解析して共起データを抽出し、さらに抽出した共起データに対して関連語辞書を利用した補完を行う。

【００１１】また、抽出された共起データを補完する方法が、特公平５－３４６９３８号公報「共起データの補完方式」に開示されている。この方法によると、抽出された共起データに関する統計情報を利用して第１および第２の構成単語間の意味的な距離を算出し、第１の構成単語と第２の構成単語の組合せのうち、もとの構成単語と意味的に近いものを共起データとして補完する。

【００１２】上記２つの方法は、文書から共起データを抽出する方法について開示しているが、抽出した共起データをどの分類軸に属す概念であるか推定する方法や、抽出した共起データ同士に共通する概念見出しを付与する方法については述べていない。

【００１３】

【発明が解決しようとする課題】以上述べたように、従来、文書分類に必要な辞書はあらかじめ用意されているものであり、ユーザが辞書を更新する手段は提供されていなかった。

【００１４】本発明は、この点に鑑み、あらかじめ設定された複数の分類軸に対して、各分類軸に属す概念と該概念を表す表現とを対応づけた概念定義辞書を利用して、各文書中に含まれる概念を抽出し、抽出された概念のうち、異なる分類軸に属す概念同士を組み合わせた複合概念を用いて文書を分類するシステムにおいて、辞書の更新が必要かどうかを判断し、更新の必要がある場合には辞書登録すべき候補の一覧を提示することにより、辞書のメンテナンスが容易な文書分類システムを提供することを目的とする。

【００１５】

【課題を解決するための手段】そこで、本願発明の文書情報抽出方法は、文書を入力する入力部と、前記入力部により入力された文書を解析し、あらかじめ登録した共起パターンに適合する表現を、少なくとも、第１の構成要素、第２の構成要素、要素間の関係、第２の構成要素に付随する付属語列とともに記憶する共起抽出部と、前記共起抽出部により抽出された共起データに関して、要素間の関係または第２の構成要素に付随する付属語列を参照して分類する共起分類部、とを具備することを特徴とする（請求項１）。

【００１６】このような構成によれば、与えられた文書から分類軸に関するとともに、共起データを抽出することができる。

【００１７】このとき、同義語、類義語、シソーラス等の関連語を記憶した関連語辞書を用意し、抽出された各共起データの構成要素を関連語辞書に記憶された関連語に置き換え、各共起データをより上位の概念へと変換し、共起データ、分類軸、変換された上位概念とを対応づけて記憶することで、各共起データの上位概念を抽出することができる（請求項２）。

【００１８】また、本発明の文書分類システムは、複数の文書を入力する入力手段と、あらかじめ設定された複数の分類軸に対して、各分類軸に属す概念と該概念を表す表現とを対応づけた概念定義辞書と、前記概念定義辞書を用いて、前記入力手段により入力された各文書中に含まれる概念を抽出する概念抽出手段と、前記概念抽出手段により抽出された概念のうち、異なる分類軸に属す概念同士を組み合わせた複合概念を用いて前記入力された文書を分類する文書分類手段と、前記文書分類手段の分類性能を評価する分類性能評価手段と、前記分類性能評価手段の評価結果に基づき、前記概念定義辞書への辞書登録候補を自動生成する辞書作成手段と、前記辞書作成手段により作成された辞書登録候補を表示する表示手段、とを具備することを特徴とする（請求項３）。

【００１９】このような構成によれば、分類性能を評価することで、辞書の更新が必要かどうかを判断し、更新の必要がある場合には辞書登録すべき候補の一覧を提示することができる。

【００２０】このとき、各分類に属す文書件数とあらかじめ設定した値とを比較したり（請求項４）、入力された文書件数といずれかの分類に属す文書件数とあらかじめ設定した値とを比較することで（請求項５）、分類性能評価を行うことができる。

【００２１】また、上記辞書作成手段は、入力手段により入力された文書を解析し、あらかじめ登録した共起パターンに適合する表現を、少なくとも、第１の構成要素、第２の構成要素、要素間の関係、第２の構成要素に付随する付属語列とともに記憶する共起抽出手段と、共起抽出手段により抽出された共起データに関して、要素間の関係または第２の構成要素に付随する付属語列を参照して、各共起データが属す分類軸を推定する分類軸推定手段と、同義語、類義語、シソーラス等の関連語を記憶した関連語辞書と、共起抽出手段により抽出された各共起データの構成要素を関連語辞書に記憶された関連語に置き換え、各共起データをより上位の概念へと変換する概念変換手段とを有し、共起データ、分類軸、上位概念を対応づけて記憶することを特徴とする（請求項６）。

【００２２】このような構成によれば、分類軸や概念見出とともに、辞書登録すべき候補の一覧を提示することができる。

【００２３】

【発明の実施の形態】［第一の実施形態］以下、図面を参照して本発明に係る第一の実施形態を説明する。図１は本発明の第一の実施形態に係る文書分類システムの構成を示すブロック図である。なお、本実施形態における文書分類システムは、例えば磁気ディスク等の記録媒体に記録されたプログラムを読み込み、このプログラムによって動作が制御されるコンピュータによって実現される。

【００２４】図１において、入力手段としての入力部１０１は、処理対象となる文書を例えばメモリや磁気ディスク、光ディスクなどから取り込む。

【００２５】制御部１０２は入力部１０１が取り込んだ情報を受け取り解析した後、各処理部へ必要な情報を送る。各処理部での処理結果は再び制御部１０２に返され、必要な情報は出力部１０３を介して出力され、例えばディスプレイに表示される。

【００２６】処理結果保持部１１０は、各処理部の結果を一時的に保持するための記憶領域であり、例えばＲＡＭや磁気ディスクなどからなる。この処理結果保持部１１０には、形態素解析結果リスト１１０ａ、共起リスト１１０ｂが設けられている。各リストに記憶される情報については、各処理部の説明の中で詳述する。

【００２７】まず、制御部１０２は、入力部１０１から取り込まれた文書を共起抽出部１０４に送る。共起抽出部１０４は、形態素解析部１０７を起動する。

【００２８】形態素解析部１０７は形態素解析辞書１０９を参照しながら、形態素解析を行い、その結果を処理結果保持部１１０に設けられた形態素解析結果リスト１１０ａに記憶する。形態素解析部１０７の処理動作については広く公知であるので、説明を省略する。

【００２９】つぎに、共起抽出部１０４は、形態素解析部１０７から受け取った形態素解析結果リスト１１０ａから、共起パターン表１０８に記憶されるパターンと一致する表現を抽出し、その結果を処理結果保持部１１０に設けられた共起リスト１１０ｂに記憶する。

【００３０】つぎに、制御部１０２は、共起抽出部１０４の処理結果である共起リスト１１０ｂを共起分類部１０５に送る。共起分類部１０５は、共起リスト１１０ｂに記憶される情報のうち、要素間の関係あるいは第２の構成要素に付随する付属語列を参照して、各共起データが属す分類軸を推定し、その結果を共起リスト１１０ｂに記憶する。

【００３１】つぎに、制御部１０２は、共起分類部１０５の処理結果である共起リスト１１０ｂを概念変換部１０６に送る。概念変換部１０６は、共起リスト１１０ｂに記憶される各共起データの構成要素を関連語辞書１１１に記憶される関連語に置き換え、各共起データをより上位の概念へと変換して概念見出しを作成し、その結果を共起リスト１１０ｂに記憶する。

【００３２】以上が本システムの概要である。次に、各処理部の詳細についてフローチャートを用いて説明する。

（ａ）共起抽出部１０４の処理動作

共起抽出部１０４の処理動作について説明する。図２は共起抽出部１０４の処理動作を示すフローチャートである。

【００３３】まず、ステップＳ２０１で共起リスト１１０ｂを初期化し、ステップＳ２０２に進む。ステップＳ２０２で入力文書を読み込み、ステップＳ２０３に進む。ステップＳ２０３で、処理中の文番号を示す変数ｉに初期値１をセットし、ステップＳ２０４に進む。

【００３４】ステップＳ２０４で、未処理の文が残っているかどうか判定する。残っている場合はステップＳ２０５に進み、残っていない場合は、処理を終了する。

【００３５】ステップＳ２０５で、ｉ番目の文を形態素解析部１０７に送り、その処理結果を受け取り、ステップＳ２０６に進む。

【００３６】ステップＳ２０６で、処理中の文節番号を示す変数ｊに初期値１をセットし、ステップＳ２０７に進む。

【００３７】ステップＳ２０７でｊ＋１＜＝文節数であるかどうか判定する。ｊ＋１が文節数に等しいか小さい場合にはステップＳ２０８に進み、そうでない場合はステップＳ２１３に進む。

【００３８】ステップＳ２０８で、文節ｊ、文節ｊ＋１に一致する共起パターンが共起パターン表１０８にあるかどうか判定する。一致するパターンがある場合にはステップＳ２０９に進み、ない場合にはステップＳ２１２に進む。

【００３９】ステップＳ２０９で、文節ｊ、文節ｊ＋１からなる共起データが共起リスト１１０ｂに登録済かど

うか判定する。すでに登録されている場合はステップＳ２１１に進み、まだ登録されていない場合はステップＳ２１０に進む。

【００４０】ステップＳ２１０で、文節ｊ、文節ｊ＋１からなる共起データを共起リスト１１０ｂに登録し、ステップＳ２１１に進む。ステップＳ２１１で、該当する共起データの頻度に１を加算し、ステップＳ２１２に進む。

【００４１】ステップＳ２１２でｊを１インクリメントし、ステップＳ２０７に戻る。ステップＳ２０７からステップＳ２１２までの繰り返しにより、ｉ番目の文に存在する共起データの抽出を終了すると、ステップＳ２１３に進む。ステップＳ２１３で、ｉを１インクリメントし、ステップＳ２０４に戻る。ステップＳ２０４からステップＳ２１３の繰り返しにより、すべての文に存在する共起データの抽出を終了する。

【００４２】図３に共起パターン表１０８に記憶される情報の例を示す。共起パターン表１０８は第１要素、第１要素と第２要素の関係、第２要素から構成されている。第１要素と第２要素には品詞情報が、関係には付属語情報が格納されている。付属語を伴わない関係の場合には、関係の欄は空になる。

【００４３】図４に共起リスト１１０ｂに記憶される情報の例を示す。共起リスト１１０ｂは第１要素、第１要素と第２要素の関係、第２要素、第２要素に付随する付属語列、頻度、分類軸、概念見出しから構成されている。共起抽出部１０４の処理が終了した段階では、図４（ａ）に示すように、分類軸と概念見出しの欄は空である。

【００４４】このようにして、共起抽出部１０５は、形態素解析部１０７を起動して形態素解析を行い、その結果を共起パターン表１０８に記憶される各パターンと照合して一致する情報を抽出し、その結果を共起リスト１１０ｂに記憶する。

【００４５】（ｂ）共起分類部１０５の処理動作

つぎに、共起分類部１０５の処理動作について説明する。図５は共起分類部１０５の処理動作を示すフローチャートである。

【００４６】まず、ステップＳ５０１で、共起リスト１１０ｂを読み込み、ステップＳ５０２に進む。ステップＳ５０２で、処理中のリスト番号を示す変数ｉに初期値１をセットし、ステップＳ５０３に進む。ステップＳ５０３でｉ＜＝リスト件数であるかどうか判定する。ｉがリスト件数に等しいか小さい場合はステップＳ５０４に進み、そうでない場合は処理を終了する。

【００４７】ステップＳ５０４で、ｉ番目のリストの第２の構成要素が動詞であるかどうか判定する。動詞である場合はステップＳ５０５に進み、そうでない場合はステップＳ５０８に進む。

【００４８】ステップＳ５０５で、ｉ番目のリストの要素間の関係が「を」であるかどうか判定する。「を」である場合はステップＳ５０９に進み、そうでない場合はステップＳ５０６に進む。

【００４９】ステップＳ５０６で、ｉ番目のリストの第２の構成要素に付随する付属語列が「ている」であるかどうか判定する。「ている」である場合はステップＳ５１０に進み、そうでない場合はステップＳ５０７に進む。

【００５０】ステップＳ５０７で、ｉ番目のリストの第２の構成要素が動詞「ある」であるかどうか判定する。「ある」である場合には、ステップＳ５１０に進み、そうでない場合はステップＳ５０９に進む。

【００５１】ステップＳ５０８で、ｉ番目のリストの第２の構成要素が形容詞または形容動詞であるかどうか判定する。形容詞または形容動詞である場合はステップＳ５１０に進み、そうでない場合はステップＳ５１１に進む。

【００５２】ステップＳ５０９で、分類軸を「アクション」とし、ステップＳ５１２に進む。ステップＳ５１０で、分類軸を「状態」とし、ステップＳ５１２に進む。ステップＳ５１１で、分類軸を「体言」とし、ステップＳ５１２に進む。

【００５３】ステップＳ５１２で、分類軸を共起リスト１１０ｂに追加し、ステップＳ５１３に進む。

【００５４】ステップＳ５１３で、ｉを１インクリメントし、ステップＳ５０３に戻る。ステップＳ５０３からステップＳ５１３までの繰り返しにより、共起リスト１１０ｂに記憶されるすべての共起データについて、分類軸がセットされ、処理を終了する。

【００５５】共起分類部１０５の処理が終了した段階で、共起リスト１１０ｂには、図４（ｂ）に示すように分類軸が書き込まれる。

【００５６】このようにして、共起分類部１０５は、共起リスト１１０ｂに記憶される情報のうち、要素間の関係あるいは第２の構成要素に付随する付属語列を参照して、各共起データが属す分類軸を推定し、その結果を共起リスト１１０ｂに記憶する。

【００５７】（ｃ）概念変換部１０６の処理動作

つぎに、概念変換部１０６の処理動作について説明する。図６は概念変換部１０６の処理動作を示すフローチャートである。

【００５８】まず、ステップＳ６０１で、共起リスト１１０ｂを読み込み、ステップＳ６０２に進む。ステップＳ６０２で、処理中のリスト番号を示す変数ｉに初期値１をセットし、ステップＳ６０３に進む。ステップＳ６０３でｉ＜＝リスト件数であるかどうか判定する。ｉがリスト件数に等しいか小さい場合はステップＳ６０４に進み、そうでない場合は処理を終了する。

【００５９】ステップＳ６０４で、ｉ番目の共起データに関して、第１の構成要素をＷＯＲＤ１に、第２の構成

要素をＷＯＲＤ２に、要素間の関係をＦＺＫに、各々セットし、ステップＳ６０５に進む。

【００６０】ステップＳ６０５で、関連語辞書１１１を参照して、ＷＯＲＤ１の関連語があるかどうか判定する。関連語がある場合は、ステップＳ６０６に進み、ない場合はステップＳ６０７に進む。ステップＳ６０６で、該当する関連語をＷＯＲＤ１にセットし、ステップＳ６０７に進む。

【００６１】ステップＳ６０７で、関連語辞書１１１を参照して、ＷＯＲＤ２の関連語があるかどうか判定する。関連語がある場合は、ステップＳ６０８に進み、ない場合はステップＳ６０９に進む。

【００６２】ステップＳ６０８で、該当する関連語をＷＯＲＤ２にセットし、ステップＳ６０９に進む。

【００６３】ステップＳ６０９で、関連語辞書１１１を参照して、ＦＺＫの関連語があるかどうか判定する。関連語がある場合は、ステップＳ６１０に進み、ない場合はステップＳ６１１に進む。ステップＳ６１０で、該当する関連語をＦＺＫにセットし、ステップＳ６１１に進む。

【００６４】ステップＳ６１１で、ＷＯＲＤ１、ＦＺＫ、　ＷＯＲＤ２を概念見出しとして共起リスト１１０ｂに記憶し、ステップＳ６１２に進む。ステップＳ６１２でｉを１インクリメントし、ステップＳ６０３に戻る。ステップＳ６０３でからステップＳ６１２までの繰り返しにより、共起リスト１１０ｂに記憶されるすべての共起データについて、概念見出しがセットされ、処理を終了する。

【００６５】図７に、関係語辞書１１１に記憶される情報の例を示す。関係語辞書１１１は、表現とその上位概念から構成される。関連語辞書１１１には、自立語だけでなく付属語を記憶してもよい。

【００６６】概念変換部１０６の処理が終了した段階で、共起リスト１１０ｂには、図４（ｃ）に示すように概念見出しが書き込まれる。

【００６７】このようにして、概念変換部１０６は、共起リスト１１０ｂに記憶される各共起データの構成要素を関連語辞書１１１に記憶される関連語に置き換え、各共起データをより上位の概念へと変換して概念見出しを作成し、その結果を共起リスト１１０ｂに記憶する。

【００６８】以上のようにして、与えられた文書から、分類軸と概念見出しを伴った共起データを抽出することができる。

【００６９】［第二の実施形態］次に、本発明に係る第二の実施形態を説明する。図８は本発明の第二の実施形態に係る文書分類システムの構成を示すブロック図である。第二の実施形態は、第一の実施形態で説明した文書情報抽出方法を、文書分類システムに応用した例である。

【００７０】図８において、図１と同一部分には同一符号を付し、異なる部分についてのみ説明する。すなわち、図８の制御部１０２には、文書分類制御部８０１、辞書更新制御部８０２、処理結果保持部１１０が接続している。文書分類制御部８０１には、概念抽出部８０３、文書分類部８０５、分類性能評価部８０６、概念定義辞書８０４が接続している。一方、辞書更新制御部８０２には、共起抽出部１０４、共起分類部１０５、概念変換部１０６、共起パターン表１０８、形態素解析部１０７、形態素解析辞書１０９、関連語辞書１１１が接続している。

【００７１】辞書更新制御部８０２以下の処理動作は、図１における制御部１０２以下の処理動作と同様であるので、説明を省略する。

【００７２】処理結果保持部１１０には、抽出概念リスト１１０ｃ、分類リスト１１０ｄ、入力文書数１１０ｅ、辞書候補リスト１１０ｆが追加されている。

【００７３】図８において、入力手段としての入力部１０１は、例えばキーボード、マウス、ペン入力装置などからなり、文字列の入力や選択操作指示などのコマンド入力を取り込む。また、処理対象となる文書を例えばメモリや磁気ディスク、光ディスクなどから取り込む。

【００７４】制御部１０２は入力部１０１が取り込んだ情報を受け取り解析した後、各処理部へ必要な情報を送る。各処理部での処理結果は再び制御部１０２に返され、必要な情報は出力部１０３を介して出力され、例えばディスプレイに表示される。

（ｄ）制御部１０２の処理動作

まず、制御部１０２の処理動作について説明し、全体の処理の流れを述べる。図９は制御部１０２の処理動作を示すフローチャートである。

【００７５】まず、ステップＳ９０１で、文書分類制御部８０１を起動し、文書分類制御部８０１に付随する各種処理を実行させ、ステップＳ９０２に進む。ステップＳ９０２で、分類性能評価部８０６の評価結果を受け取り、ステップＳ９０３に進む。ステップＳ９０３で、評価結果がであるかどうか判定する。である、すなわち分類性能が良いと判定された場合は、ステップＳ９０５に進む。でない、すなわち分類性能が悪いと判定された場合は、ステップＳ９０４に進む。

【００７６】ステップＳ９０４で、辞書更新制御部８０２を起動し、辞書更新制御部８０２に付随する各種処理を実行させ、ステップＳ９０６に進む。ステップＳ９０６で辞書候補１１０ｆリストを受け取り、ステップＳ９０７に進む。ステップＳ９０７で受け取った辞書候補リスト１１０ｆを出力部１０３に送り、処理を終了する。

【００７７】一方、ステップＳ９０３からステップＳ９０５に進んだ場合は、ステップＳ９０５で、辞書更新不要のメッセージを出力部１０３に送り、処理を終了する。

【００７８】以上が本システムの概要である。次に、各

処理部の詳細についてフローチャートを用いて説明する。

（ｅ）概念抽出部８０３の処理動作

つぎに、概念抽出部８０３の処理動作について説明する。図１０は概念抽出部８０３の処理動作を示すフローチャートである。

【００７９】まず、ステップＳ１００１で、抽出概念リスト１１０ｃを初期化し、ステップＳ１００２に進む。ステップＳ１００２で概念定義辞書８０４を読み込み、ステップＳ１００３に進む。

【００８０】この処理により、概念定義辞書８０４のうち処理に必要な情報が、変数ＤＩＣ＿ＩＤ、ＤＩＣ＿ＨＹＯＧＥＮ、ＤＩＣ＿ＣＮＴ　にセットされる。すなわち、概念定義辞書８０４のｉ行目の概念ＩＤがＤＩＣ＿ＩＤ［ｉ］に、表現がＤＩＣ＿ＨＹＯＧＥＮ［ｉ］に、概念定義辞書８０４の行数がＤＩＣ＿ＣＮＴ　にセットされる。

【００８１】ここで、図１５に概念定義辞書８０４に記憶される情報の一例を示す。概念ＩＤを示す記号、分類軸、概念見出し、概念を表す表現から構成される。概念を表す表現は、例えば形態素解析や構文解析の処理結果にアクセスする形で記述してもよい。

【００８２】図１０の説明に戻り、ステップＳ１００３で処理中の文書を示す変数ｉに初期値１をセットし、ステップＳ１００４に進む。ステップＳ１００４で、未処理の文書が残っているかどうか判定する。残っている場合はステップＳ１００５に進み、残っていない場合はステップＳ１０１２に進む。

【００８３】ステップＳ１００５でｉ番目の文書を読み込み、抽出概念リスト１１０ｃに文書ＩＤを追加し、ステップＳ１００６に進む。

【００８４】ステップＳ１００６で処理中の概念定義辞書８０４の行を示す変数ｊに初期値１をセットし、ステップＳ１００７に進む。ステップＳ１００７でｊ＜＝ＤＩＣ＿ＣＮＴ　であるかどうか判定する。ｊがＤＩＣ＿ＣＮＴ　に等しいか小さい場合は、ステップＳ１００８に進む。そうでない場合はステップＳ１０１１に進む。

【００８５】ステップＳ１００８で、ｉ番目の文書がＤＩＣ＿ＨＹＯＧＥＮ［ｊ］に適合する部分文字列を含んでいるかどうか判定する。適合する部分文字列を含んでいる場合はステップＳ１００９に進み、そうでない場合はステップＳ１０１０に進む。ステップＳ１００９で抽出概念リスト１１０ｃにＤＩＣ＿ＩＤ［ｊ］を追加する。

【００８６】ステップＳ１０１０でｊを１インクリメントし、ステップＳ１００７に戻る。ステップＳ１００７からステップＳ１０１０の繰り返しにより、概念定義辞書８０４のすべての行に対する照合を終了すると、ステップＳ１０１１に進む。ステップＳ１０１１でｉを１インクリメントし、次の文書の処理に移る。

【００８７】ステップＳ１０１２で、入力文書数１１０ｅにｉ－１を記憶し、処理を終了する。

【００８８】図１６に抽出概念リスト１１０ｃに記憶される情報の例を示す。抽出概念リスト１１０ｃは文書ＩＤ、抽出された概念ＩＤから構成される。１文書内に複数の概念が抽出された場合には、カンマで区切って概念ＩＤが記述される。

【００８９】以上のようにして、概念抽出部８０３は、概念定義辞書８０４に記憶される情報を用いて、与えられた文書に含まれる概念を抽出し、その結果を抽出概念リスト１１０ｃに記憶する。また、入力された文書の数を入力文書数１１０ｅに記憶する。

【００９０】（ｆ）文書分類部８０５の処理動作

つぎに、文書分類部８０５の処理動作について説明する。図１１は文書分類部８０５の処理動作を示すフローチャートである。

【００９１】まず、ステップＳ１１０１で、分類リスト１１０ｄを初期化し、ステップＳ１１０２に進む。ステップＳ１１０２で抽出概念リスト１１０ｃと入力文書数１１０ｅを読み込み、ステップＳ１１０３に進む。

【００９２】ステップＳ１１０３で概念定義辞書８０４を読み込み、ステップＳ１１０４に進む。この処理により、概念定義辞書８０４の内容が、変数ＤＩＣ＿ＩＤ、ＤＩＣ＿ＪＩＫＵ、　ＤＩＣ＿ＣＮＴ　にセットされる。

【００９３】ステップＳ１１０４で分類項目を作成し、ステップＳ１１０５に進む。この処理により、分類すべき項目一覧が、変数ｂｕｎｒｕｉ＿ｉｄ　、ｂｕｎｒｕｉ＿ｃｎｔにセットされる。このステップＳ１１０４の処理動作については後述する。

【００９４】ステップＳ１１０５で処理中の文書を示す変数ｉに初期値１をセットし、ステップＳ１１０６に進む。ステップＳ１１０６でｉ＜＝入力文書数であるかどうか判定する。ｉが入力文書数に等しいか小さい場合は、ステップＳ１１０７に進み、そうでない場合は処理を終了する。

【００９５】ステップＳ１１０７で処理中の分類項目を示す変数ｊに初期値１をセットし、ステップＳ１１０８に進む。ステップＳ１１０８でｊ＜＝ｂｕｎｒｕｉ＿ｃｎｔであるかどうか判定する。ｊがｂｕｎｒｕｉ＿ｃｎｔに等しいか小さい場合は、ステップＳ１１０９に進み、そうでない場合はステップＳ１１１４に進む。

【００９６】ステップＳ１１０９で、ｉ番目の文書はｂｕｎｒｕｉ＿ｉｄ［ｊ］に含まれる概念のすべてを含むかどうか判定する。ｂｕｎｒｕｉ＿ｉｄ［ｊ］に含まれるすべての概念を含む場合にはステップＳ１１１０に進み、そうでない場合はステップＳ１１１３に進む。

【００９７】ステップＳ１１１０で分類リスト１１０ｄにｂｕｎｒｕｉ＿ｉｄ［ｊ］が存在するかどうか判定する。存在しない場合はステップＳ１１１１に進み、存在

する場合はステップＳ１１１２に進む。ステップＳ１１１１で分類リスト１１０ｄにｂｕｎｒｕｉ＿ｉｄ［ｊ］を追加し、ステップＳ１１１２に進む。ステップＳ１１１２で分類リスト１１０ｄのｂｕｎｒｕｉ＿ｉｄ［ｊ］の欄に、ｉ番目の文書の文書ＩＤを追加し、ステップＳ１１１３に進む。

【００９８】ステップＳ１１１３でｊを１インクリメントし、ステップＳ１１０８に戻る。ステップＳ１１０８からステップＳ１１１３までの繰り返しにより、すべての分類項目との照合を終了すると、ステップＳ１１１４に進む。

【００９９】ステップＳ１１１４でｉを１インクリメントし、ステップＳ１１０６に戻る。ステップＳ１１０６からステップＳ１１１４までの繰り返しにより、すべての文書のチェックを終了すると、処理を終了する。

【０１００】ここで、ステップＳ１１０４の処理動作について説明する。図１２はステップＳ１１０４の処理動作を示すフローチャートである。

【０１０１】ステップＳ１２０１で処理中の分類項目を示す変数ｍに初期値１をセットし、ステップＳ１２０２に進む。ステップＳ１２０２で変数Ｓに初期値１をセットし、ステップＳ１２０３に進む。

【０１０２】ステップＳ１２０３でＳ＜＝ＤＩＣ＿ＣＮＴ　であるかどうか判定する。ＳがＤＩＣ＿ＣＮＴ　に等しいか小さい場合はステップＳ１２０４に進み、そうでない場合はステップＳ１２１１に進む。

【０１０３】ステップＳ１２０４で変数ｔに初期値Ｓをセットし、ステップＳ１２０５に進む。ステップＳ１２０５でｔ＜＝ＤＩＣ＿ＣＮＴ　であるかどうか判定する。ｔがＤＩＣ＿ＣＮＴ　に等しいか小さい場合はステップＳ１２０６に進み、そうでない場合はステップＳ１２１０に進む。

【０１０４】ステップＳ１２０６でＤＩＣ＿ＪＩＫＵ［ｓ］！＝ＤＩＣ＿ＪＩＫＵ［ｔ］であるかどうか判定する。

【０１０５】ＤＩＣ＿ＪＩＫＵ［ｓ］　とＤＩＣ＿ＪＩＫＵ［ｔ］　が等しくない場合はステップＳ１２０７に進み、等しい場合はステップＳ１２０９に進む。

【０１０６】ステップＳ１２０７で、変数ｂｕｎｒｕｉ＿ｉｄ［ｍ］にＤＩＣ＿ＩＤ［ｓ］とＤＩＣ＿ＩＤ［ｔ］　を連結したものをセットし、ステップＳ１２０８に進む。

【０１０７】ステップＳ１２０８でｍを１インクリメントし、ステップＳ１２０９に進む。ステップＳ１２０９でｔを１インクリメントし、ステップＳ１２０５に戻る。ステップＳ１２１０でＳを１インクリメントし、ステップＳ１２０３に戻る。

【０１０８】ステップＳ１２０３からステップＳ１２１０までの繰り返しにより、異なる軸に属す概念同士の組合せを作成し、ステップＳ１２１１に進む。ステップＳ１２１１で変数ｂｕｎｒｕｉ＿ｃｎｔに分類項目数であるｍ－１をセットし、処理を終了する。ここでは、２つの軸について組合せを作成しているが、３つ以上の軸について組み合わせを作成するようにしてもよい。

【０１０９】図１７に分類リスト１１０ｄに格納される情報の例を示す。分類リスト１１０ｄは分類項目、文書ＩＤ、文書件数から構成される。１つの分類項目に対して複数の文書が該当する場合には、カンマで区切って文書ＩＤが記述される。

【０１１０】以上のようにして、文書分類部８０５は、概念抽出部８０３の処理結果を受け取り、異なった分類軸に属す概念同士を組み合わせた複合概念を用いて文書を分類し、その結果を分類リスト１１０ｄに記憶する。

【０１１１】（ｇ）分類性能評価部８０６の処理動作

つぎに、分類性能評価部８０６の処理動作について説明する。ここでは、２通りの処理動作について説明する。

【０１１２】まず、第一の処理動作について説明する。図１３は分類性能評価部８０６の第一の処理動作を示すフローチャートである。

【０１１３】まず、ステップＳ１３０１で、分類リスト１１０ｄを読み込み、ステップＳ１３０２に進む。ステップＳ１３０２で評価結果の初期値をとし、ステップＳ１３０３に進む。

【０１１４】ステップＳ１３０３で処理中のリストを示す変数ｉに初期値１をセットし、ステップＳ１３０４に進む。ステップＳ１３０４でｉ＜＝リスト件数であるかどうか判定する。ｉがリスト件数に等しいか小さい場合は、ステップＳ１３０５に進み、そうでない場合は処理を終了する。

【０１１５】ステップＳ１３０５でリストのｉ番目の分類項目に分類される文書件数＞αであるかどうされるか判定する。文書件数がαより大きい場合はステップＳ１３０６に進み、そうでない場合はステップＳ１３０７に進む。ここでαはあらかじめ設定しておく値である。

【０１１６】ステップＳ１３０６で評価結果を×とし、処理を終了する。ステップＳ１３０７でｉを１インクリメントし、ステップＳ１３０４に戻る。ステップＳ１３０４からステップＳ１３０７までの繰り返しにより、分類リスト１１０ｄ中に１つでも文書件数がαを超える分類項目があれば、評価結果は×となり、すべての分類項目についてαを超えなければ評価結果はとなる。

【０１１７】次に、第二の処理動作について説明する。図１４は分類性能評価部８０６の第二の処理動作を示すフローチャートである。

【０１１８】まず、ステップＳ１４０１で入力文書数１１０ｅと分類リスト１１０ｄを読み込み、ステップＳ１４０２に進む。

【０１１９】ステップＳ１４０２で、分類リスト１１０ｄに記述される文書ＩＤの異なり総数／入力文書数を計算し、その値がβより小さいかどうか判定する。

【０１２０】βより小さい場合はステップＳ１４０３に進み、そうでない場合はステップＳ１４０４に進む。ここでβはあらかじめ設定しておく値である。ステップＳ１４０３で評価を×、ステップＳ１４０４で評価をとし、処理を終了する。

【０１２１】以上２通りの処理動作に示したように、分類性能評価部８０６は、分類リスト１０８ｂや入力文書数１１０ｅを参照して、システムの分類性能を評価する。

【０１２２】図１９に本発明を利用したシステムの出力例を示す。図１９（ａ）は、辞書のチェックを行うかどうかの確認画面である。ここで、ユーザが実行ボタンを押すと、辞書のチェックが実行され、図１９（ｂ）または（ｃ）のように辞書チェック結果が表示される。

【０１２３】図１９（ｂ）は辞書更新の必要がないと判断された場合である。図１９（ｃ）は辞書更新の必要があると判断された場合である。この場合には、さらに辞書更新を行うかどうかの確認画面が表示される。ここで、ユーザが実行ボタンを押すと、辞書更新が実行され、図１９（ｄ）に示すように、登録すべき辞書候補一覧が表示される。ここで、実際に登録したい候補にチェックし、登録ボタンを押すと、辞書に自動登録される。

【０１２４】なお、本発明は、上記実施形態に限定されず、要旨を変更しない範囲で適宜変更して実施可能である。また、上述した実施形態において記載した方法は、コンピュータに実行させるプログラムとして、例えば、磁気ディスク（フロッピーディスク、ハードディスク等）、光ディスク（ＣＤ－ＲＯＭ、ＤＶＤ等）、半導体メモリ等の記録媒体に書き込んで各種装置に適用したり、通信媒体により伝送して各種装置に適用することも可能である。本装置を実現するコンピュータは、記録媒体に記録されたプログラムを読み込み、このプログラムによって動作が制御されることにより、上述した処理を実行する。

【０１２５】

【発明の効果】以上説明したように、上記の実施形態によれば、あらかじめ設定された複数の分類軸に対して、各分類軸に属す概念と該概念を表す表現とを対応づけた概念定義辞書を利用して、各文書中に含まれる概念を抽出し、抽出された概念のうち、異なる分類軸に属す概念同士を組み合わせた複合概念を用いて文書を分類するシステムにおいて、辞書の更新が必要かどうかを判断し、更新の必要がある場合には辞書登録すべき候補の一覧を提示することにより、辞書のメンテナンスが容易な文書分類システムを提供することができる。

【図面の簡単な説明】

【図１】本発明の第一の実施形態に係る文書情報抽出方法の概略構成を示すブロック図。

【図２】同実施形態における共起抽出部１０４の処理動作を示すフローチャート。

【図３】共起パターン表１０８に記憶される情報の一例。

【図４】共起リスト１１０ｂに記憶される情報の一例。

【図５】同実施形態における共起分類部１０５の処理動作を示すフローチャート。

【図６】同実施形態における概念変換部１０６の処理動作を示すフローチャート。

【図７】間連語辞書１１１に記憶される情報の一例。

【図８】本発明の第二の実施形態に係る文書分類システムの概略構成を示すブロック図。

【図９】同実施形態における制御部１０２の処理動作を示すフローチャート。

【図１０】同実施形態における概念抽出部８０３の処理動作を示すフローチャート。

【図１１】同実施形態における文書分類部８０５の処理動作を示すフローチャート。

【図１２】図１１におけるステップＳ１１０４の処理動作を示すフローチャート。

【図１３】同実施形態における分類性能評価部８０６の第一の処理動作を示すフローチャート。

【図１４】同実施形態における分類性能評価部８０６の第二の処理動作を示すフローチャート。

【図１５】概念定義辞書８０４に記憶される情報の一例。

【図１６】抽出概念リスト１１０ｃに記憶される情報の一例。

【図１７】分類リスト１１０ｄに記憶される情報の一例。

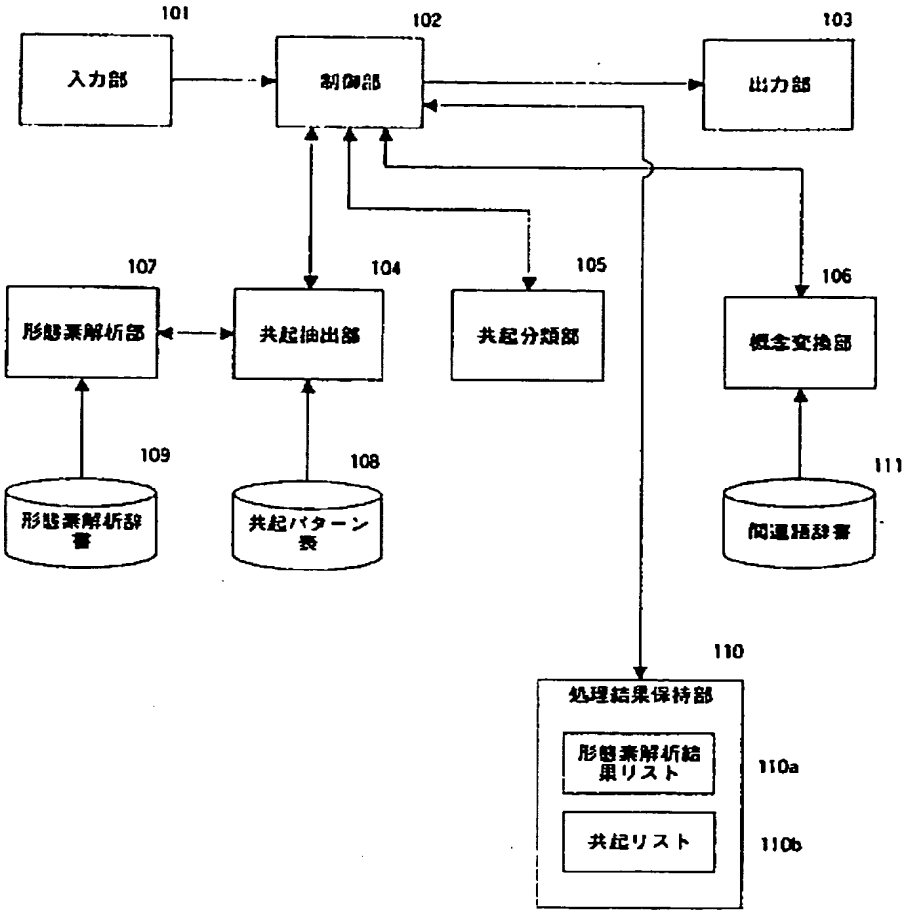
【図１８】辞書候補リスト１０８ｄに記憶される情報の一例。

【図１９】出力例。

【符号の説明】

１０１…入力部
１０２…制御部
１０３…出力部
１０４…共起抽出部
１０５…共起分類部
１０６…概念変換部
１０７…形態素解析部
１０８…共起パターン表
１０９…形態素解析辞書
１１０…処理結果保持部
８０１…文書分類制御部
８０２…辞書更新制御部
８０３…概念抽出部
８０４…概念定義辞書
８０５…文書分類部
８０６…分類性能評価部

【図１】



【図７】

表現	上位概念
良くない	悪い
売上げ	売行き

【図１７】

分類項目	文書ID	文書件数
A001+S001	#002, ……	100
A001+S002	#001, ……	200
A002+S003	#001, ……	35

【図３】

第1要素	関係	第2要素
名詞	が、は	動詞
名詞	を、も	動詞
名詞	が、は	形容詞
名詞	が、は	形容動詞

【図１５】

概念ID	分類軸	概念見出し	表現
A001	アクション	サンプルをつける	サンプルラッピング
S001	状態	売行きが悪い	売れていない
S001	状態	売行きが悪い	動きが悪い

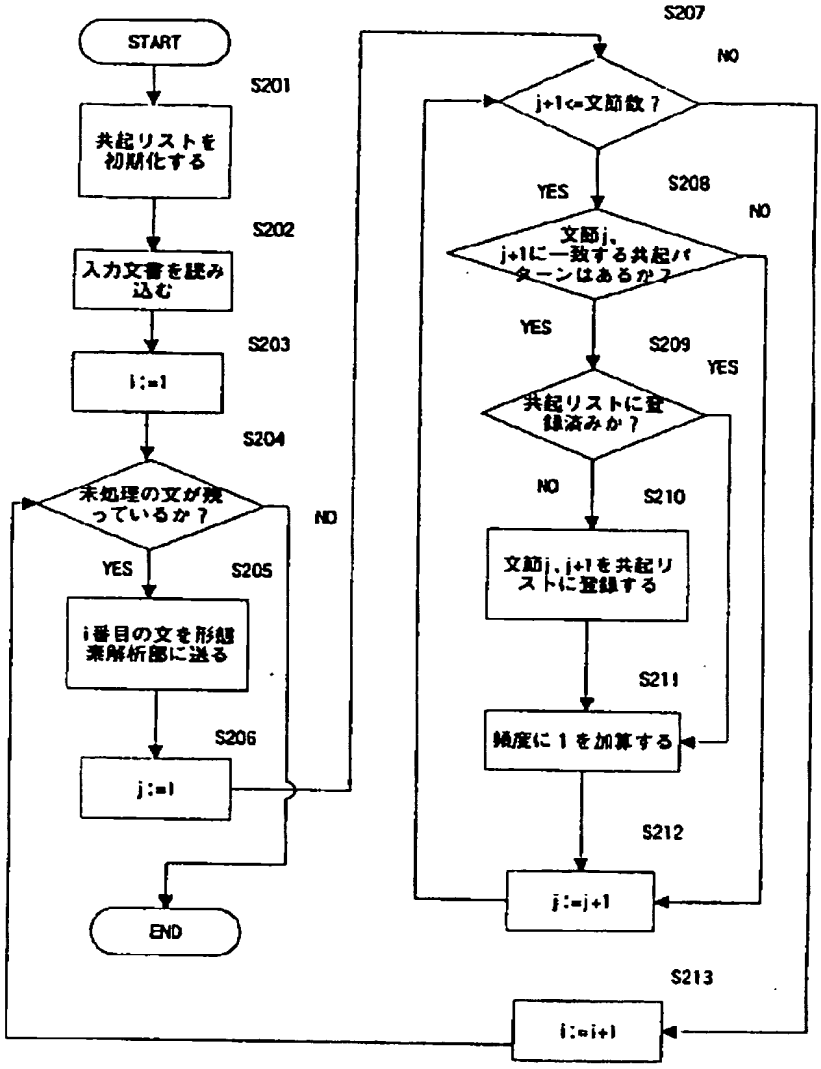
【図１６】

文書ID	概念ID
#001	A001, A002, R003, ……
#002	A001, R002, ……
#003	S005, S002, ……

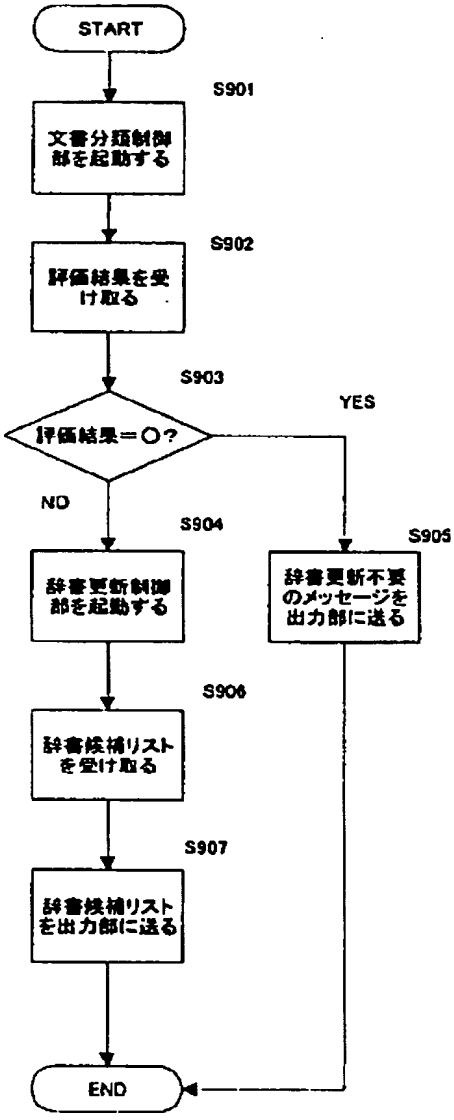
【図１８】

概念ID	分類軸	概念見出し	表現
NS001	状態	売行きが悪い	売行きが悪い
NA001	アクション	サンプルを付けた	サンプルを付けた
NS001	状態	売行きが悪い	売上げが良くない

【図２】


START
S201
共起リストを初期化する
S202
入力文書を読み込む
S203
i:=1
S204
未処理の文が残っているか？
NO
YES
S205
i番目の文を形態素解析部に送る
S206
j:=1
END
S207
NO
j+1<=文節数？
YES
S208
NO
文節j、j+1に一致する共起パターンはあるか？
YES
S209
YES
共起リストに登録済みか？
NO
S210
文節j、j+1を共起リストに登録する
S211
頻度に１を加算する
S212
j:=j+1
S213
i:=i+1

【図９】


START
S901
文書分類制御部を起動する
S902
評価結果を受け取る
S903
YES
評価結果＝○？
NO
S904
辞書更新制御部を起動する
S905
辞書更新不要のメッセージを出力部に送る
S906
辞書候補リストを受け取る
S907
辞書候補リストを出力部に送る
END

【図４】

(a)

第1要素	関係	第2要素	付属語列	頻度	分類軸	概念見出し
売行き	が	悪い		10		
サンプル	を	付け	た	5		
売上げ	が	良く	ない	20		

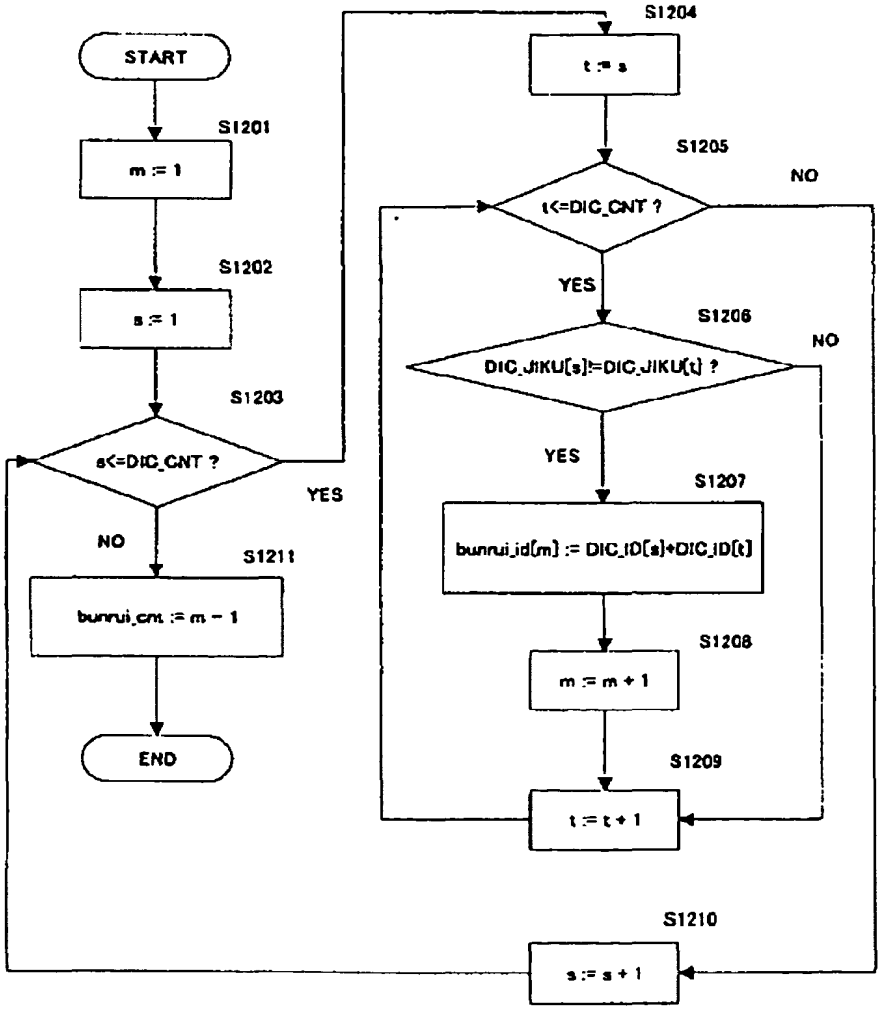
(b)

第1要素	関係	第2要素	付属語列	頻度	分類軸	概念見出し
売行き	が	悪い		10	状態	
サンプル	を	付け	た	5	アクション	
売上げ	が	良く	ない	20	状態	

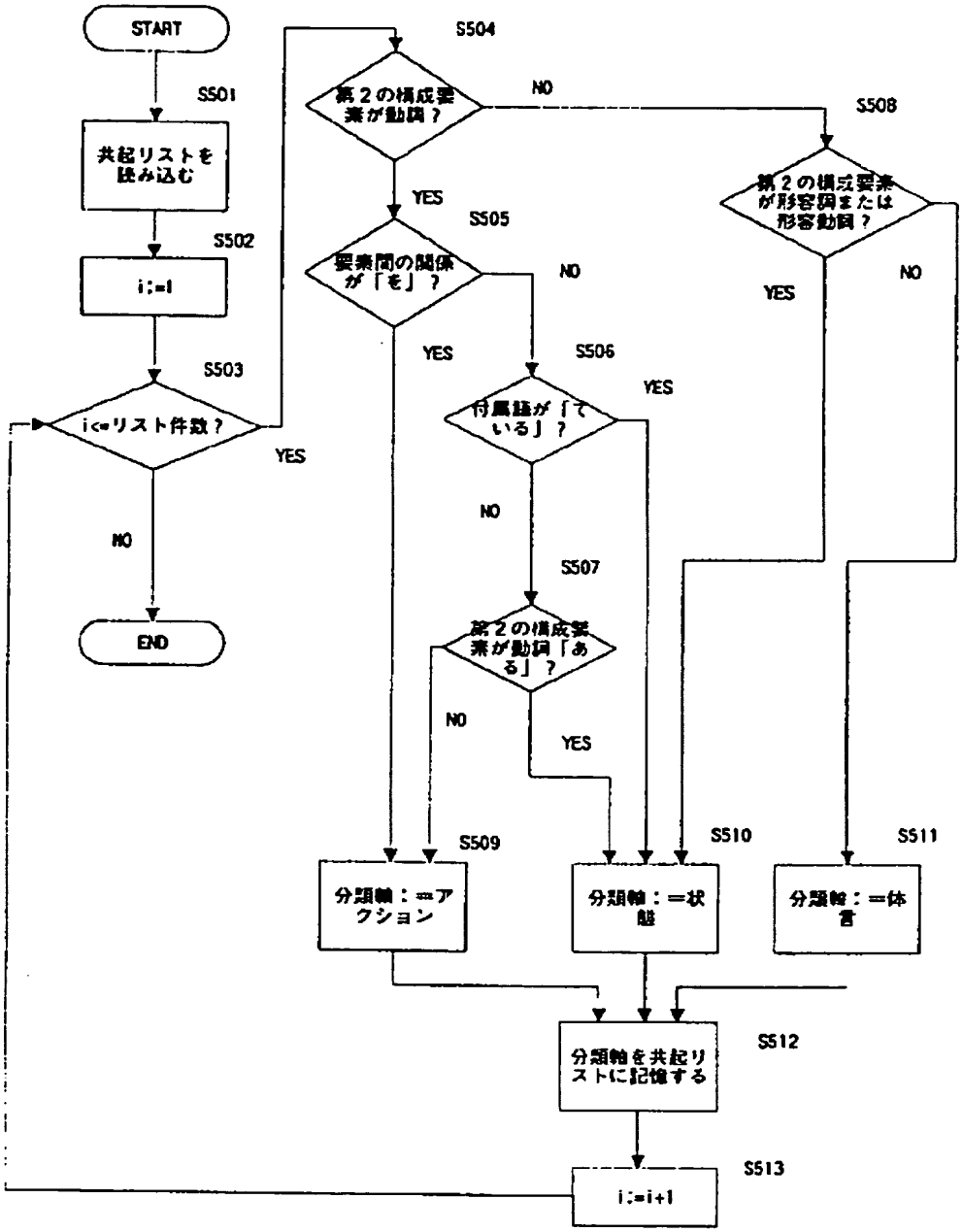
(c)

第1要素	関係	第2要素	付属語列	頻度	分類軸	概念見出し
売行き	が	悪い		10	状態	売行きが悪い
サンプル	を	付け	た	5	アクション	サンプルを付けた
売上げ	が	良く	ない	20	状態	売行きが悪い

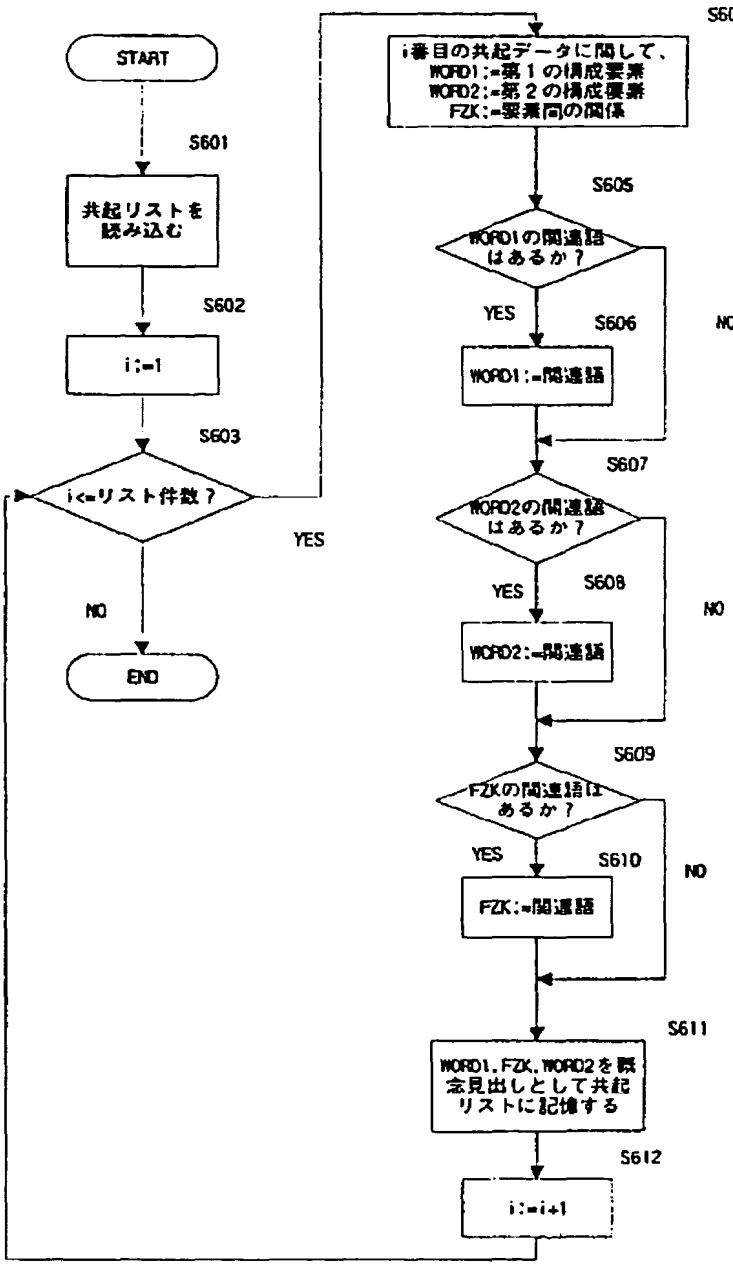
【図１２】



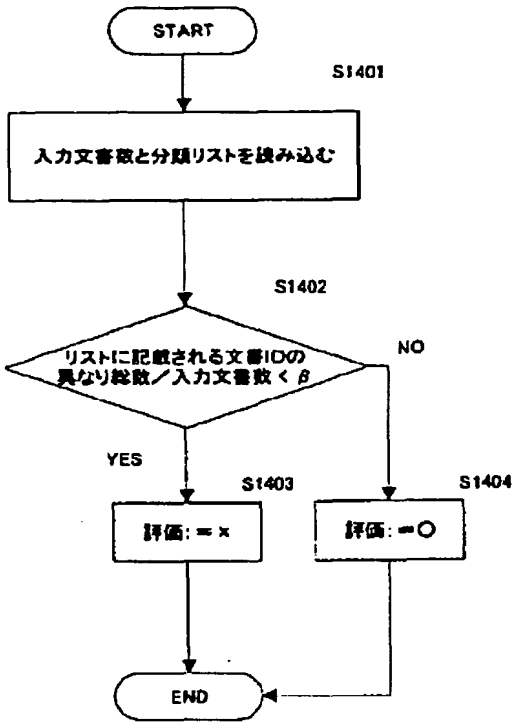
【図５】


START
S501
共起リストを読み込む
S502
i:=1
S503
i<=リスト件数？
YES
NO
END
S504
第２の構成要素が動詞？
NO
YES
S505
要素間の関係が「を」？
NO
YES
S506
付属語が「ている」？
YES
NO
S507
第２の構成要素が動詞「ある」？
NO
YES
S508
第２の構成要素が形容詞または形容動詞？
YES
NO
S509
分類軸：＝アクション
S510
分類軸：＝状態
S511
分類軸：＝体言
S512
分類軸を共起リストに記憶する
S513
i:=i+1

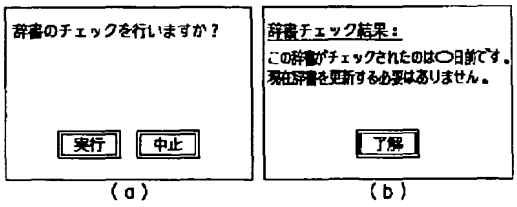
【図６】



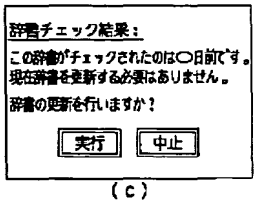
【図１４】



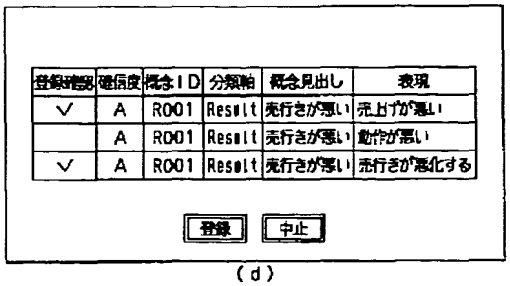
【図１９】



(a)　(b)

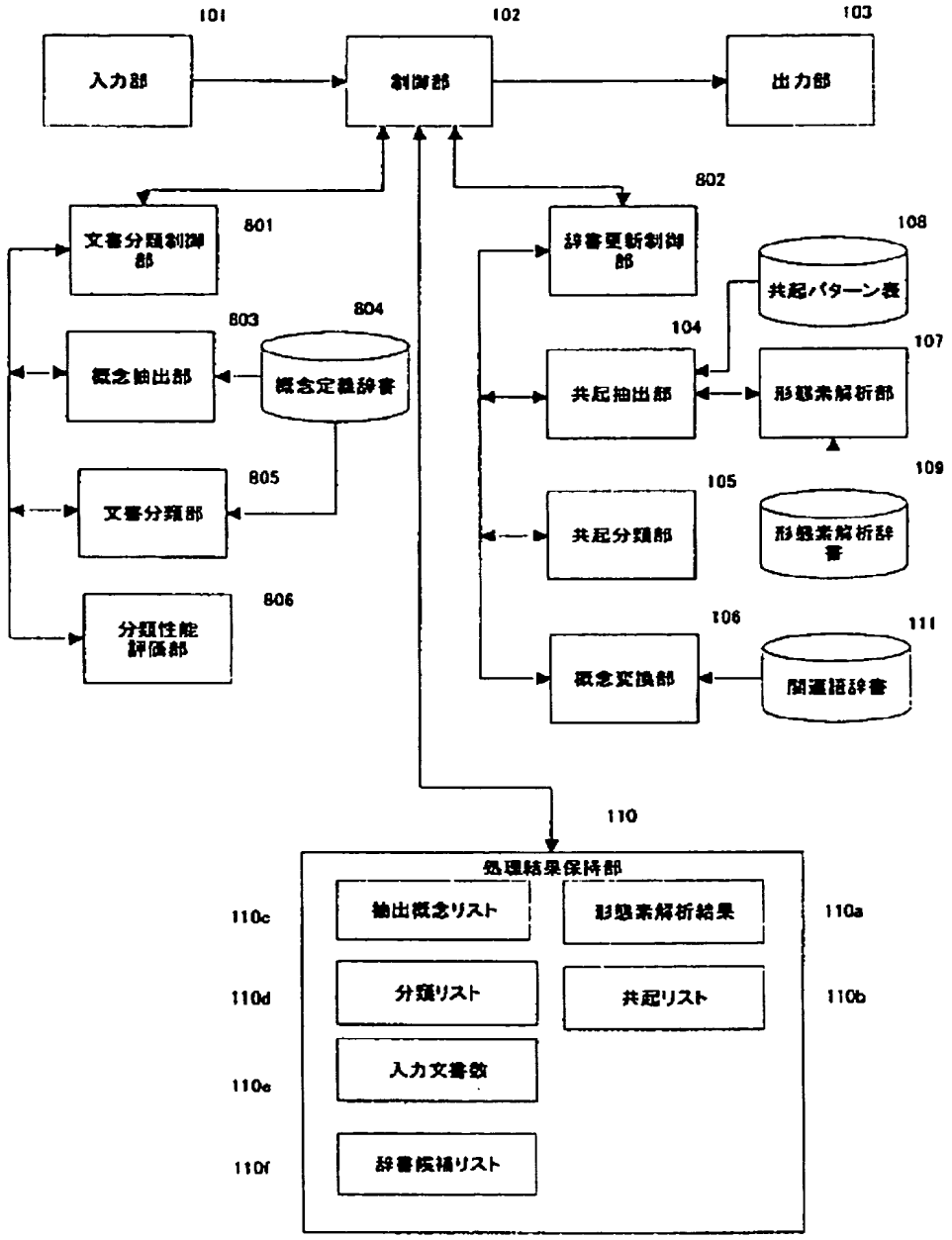


(c)

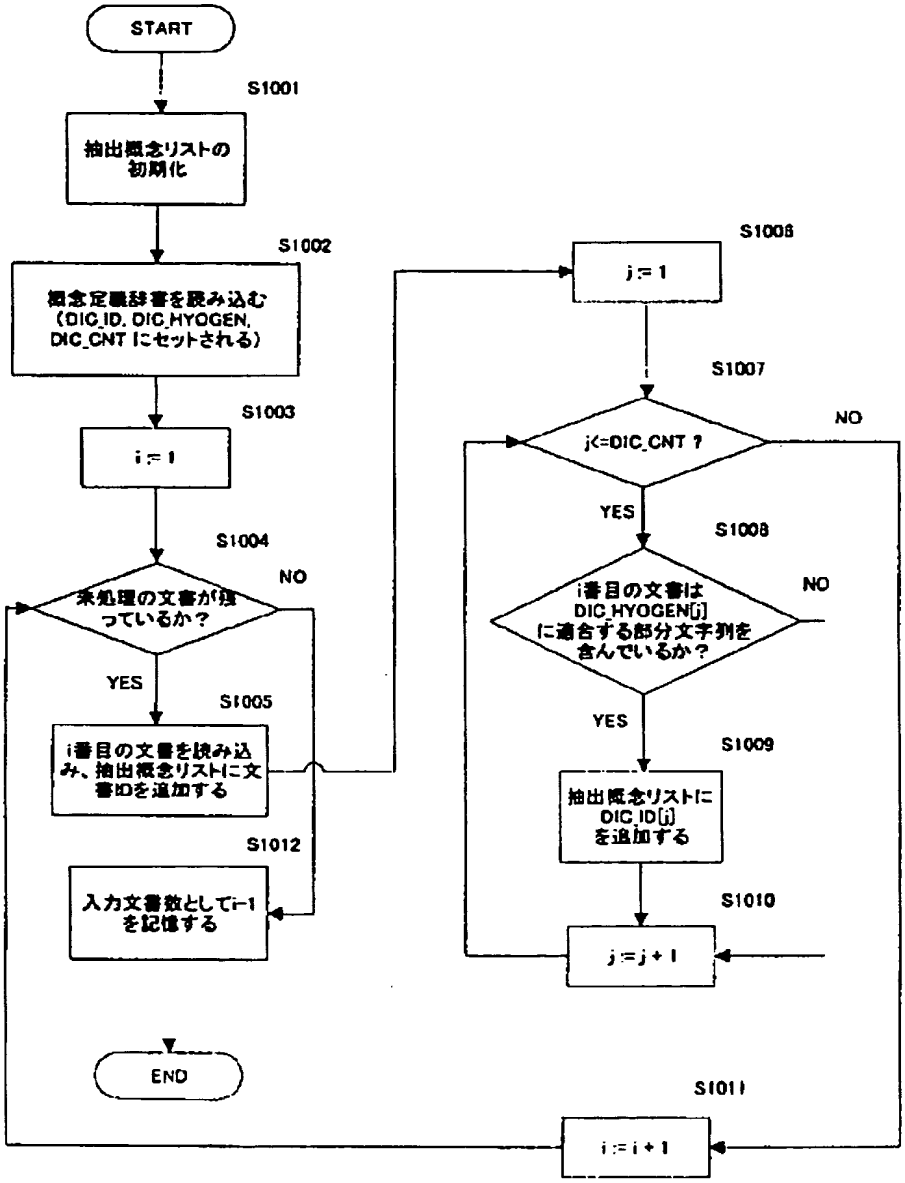


(d)

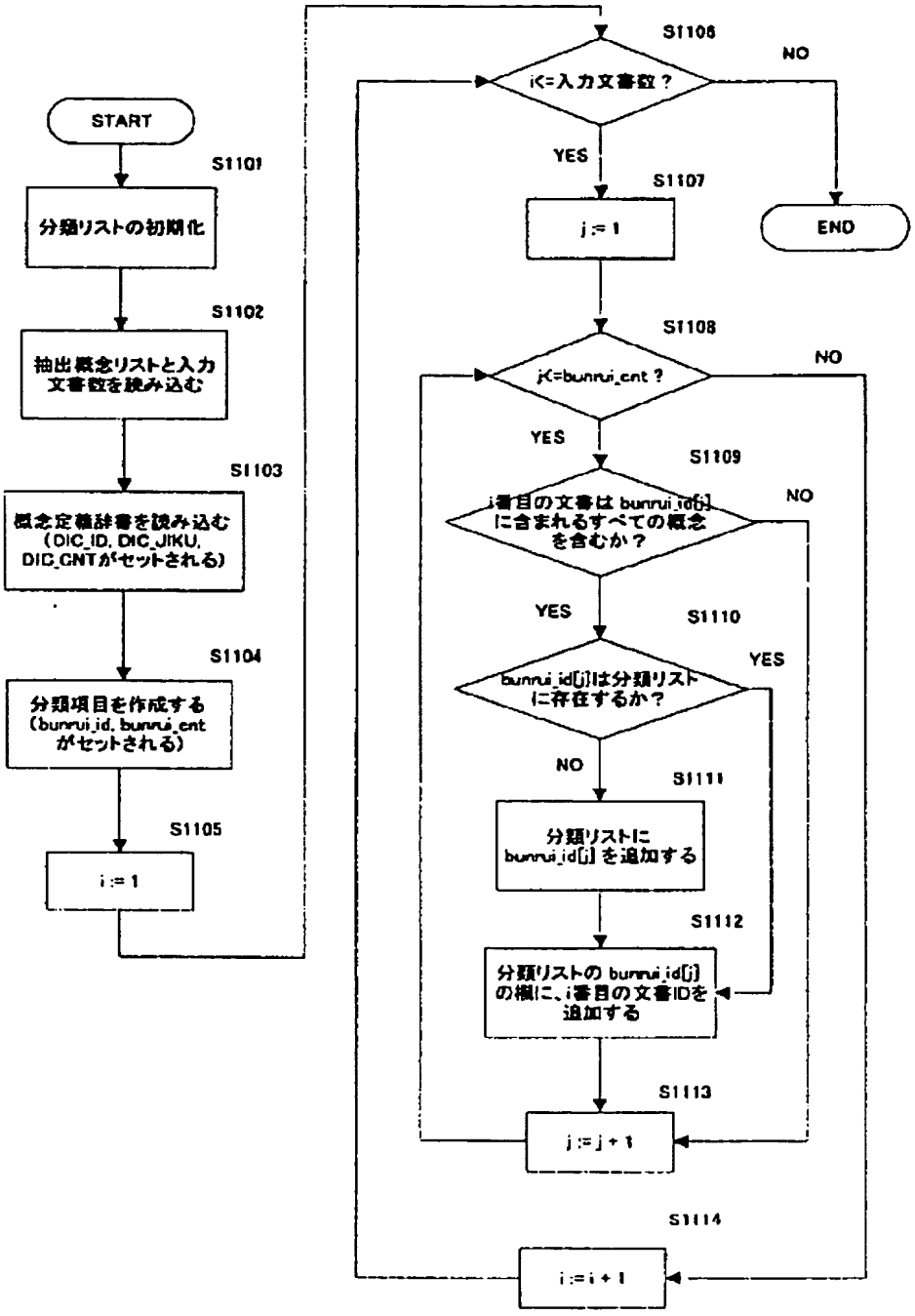
【図８】


101
入力部
102
制御部
103
出力部
801
文書分類制御部
802
辞書更新制御部
108
共起パターン表
803
804
概念抽出部
概念定義辞書
104
共起抽出部
107
形態素解析部
805
文書分類部
105
109
共起分類部
形態素解析辞書
806
分類性能評価部
106
111
概念変換部
関連語辞書
110
処理結果保持部
110c
抽出概念リスト
形態素解析結果
110a
110d
分類リスト
共起リスト
110b
110e
入力文書数
110f
辞書候補リスト

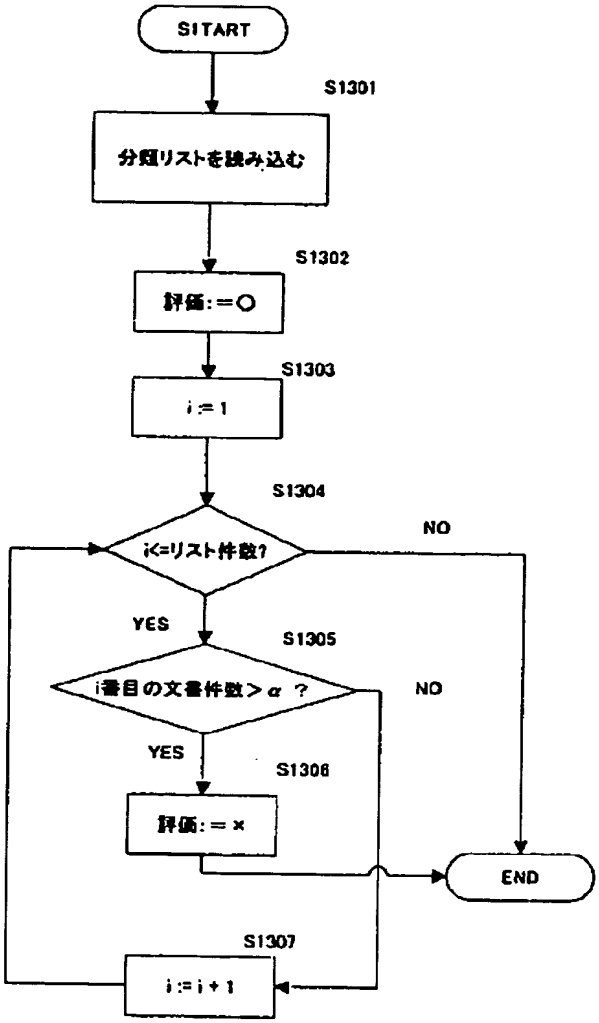
【図１０】


START
S1001
抽出概念リストの初期化
S1002
概念定義辞書を読み込む（DIC_ID, DIC_HYOGEN, DIC_CNT にセットされる）
S1003
i := 1
S1004
未処理の文書が残っているか？
NO
YES
S1005
i番目の文書を読み込み、抽出概念リストに文書IDを追加する
S1012
入力文書数としてi-1を記憶する
END
S1006
j := 1
S1007
j<=DIC_CNT ?
NO
YES
S1008
i番目の文書はDIC_HYOGEN[j]に適合する部分文字列を含んでいるか？
NO
YES
S1009
抽出概念リストにDIC_ID[j]を追加する
S1010
j := j + 1
S1011
i := i + 1

【図１１】


START
S1101
分類リストの初期化
S1102
抽出概念リストと入力文書数を読み込む
S1103
概念定義辞書を読み込む (DIC_ID, DIC_JIKU, DIC_CNTがセットされる)
S1104
分類項目を作成する (bunrui_id, bunrui_cnt がセットされる)
S1105
i := 1
S1106
i<=入力文書数？
NO
YES
END
S1107
j := 1
S1108
j<=bunrui_cnt ?
NO
YES
S1109
i番目の文書は bunrui_id[j] に含まれるすべての概念を含むか？
NO
YES
S1110
bunrui_id[j]は分類リストに存在するか？
YES
NO
S1111
分類リストに bunrui_id[j] を追加する
S1112
分類リストの bunrui_id[j] の欄に、i番目の文書IDを追加する
S1113
j := j + 1
S1114
i := i + 1

【図１３】



フロントページの続き

(51) Int. Cl.[7]	識別記号	F I	テーマコード(参考)
		G06F 15/401	310D